Development of Empirical Hydrologic and Water Quality Models of the Loxahatchee NWF Using Data-Mining Techniques

Paul Conrads
USGS- SC Water Science Center
Ed Roehl, Jr
Advanced Data Mining International, LLP

Greater Everglade Ecosystem Restoration Conference
July 13, 2010
Outline

- Data mining & Data driven models
- Modeling Loxahatchee NWR:
  - Water levels
  - Specific conductance
  - Total phosphorus
- LOXANN Decision Support System (DSS)
- DSS applications
  - Evaluation of flow releases
What is Data Mining?

- **Data Mining**: the search for valuable knowledge in massive volumes of data
- An amalgamation of techniques from various disciplines
- **Data Mining Tool Box**
  - signal processing, statistics, machine learning, chaos theory, advanced visualization
  - Artificial neural networks (ANN) models – one approach to machine learning

![Data ➔ Information ➔ Knowledge](USGS)
Data Driven Models

- Living in an era of “Big Data”
- Modeling – exercise in mapping inputs and outputs
- Empirical models - based on observations rather than on mathematically describable system processes
- Examples:
  - Linear regression: \( Y = mX + b \)
  - Artificial Neural Networks:
Loxahatchee Empirical Model

- Given inputs of flow, precipitation, and ET
- Create model(s) to simulate:
  - Water levels
  - Specific conductance
  - Total phosphorus

Inputs

Outputs
Model Architecture

SIANN = spatially interpolating artificial neural network model
Gage Height Models

Linear models based on optimal time delays and moving window averages of flow, rainfall, and ET

Spatially Interpolating ANNs – error correction models

Final prediction is the sum of the linear model and error prediction models.
**Tau Tool**

Excel application to evaluate moving window averages (MWA) and time delays of flow inputs.

### Station selection

### MWA and time delay settings

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Active</th>
<th>MinWin on/off</th>
<th>MWA on/off</th>
<th>Win</th>
<th>Tau</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RAIN2</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RAIN3</td>
<td>ON</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>RAIN4</td>
<td>OFF</td>
<td>ON</td>
<td>ON</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ET</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ET1</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ET2</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SYNC</td>
<td>ON</td>
<td>ON</td>
<td>ON</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q1</td>
<td>1-NCH</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q1c</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q2</td>
<td>2-NCH</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q2c</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q3</td>
<td>3-NCH</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q3c</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q4</td>
<td>4-NCH</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q4c</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q5</td>
<td>5-NCH</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q5c</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q6</td>
<td>6-NCH</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Q6c</td>
<td>ON</td>
<td>ON</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>LOWWE</td>
<td>OFF</td>
<td>OFF</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Statistics** – R and R²

- Low correlation between untransformed flow input and gage height: \( R^2 < 0.01 \)

**Legend**

- Blue – measured
- Red – sum Qs
- Green - error
Linear Models

Adjust MWA and time delays to increase the correlation between inputs and gage heights.

Correlation ($R^2$) increased from $<0.01$ to $>0.75$
Gage Height Error Correction Model

Model error with a spatially interpolating ANN model

Time series of linear model errors

Stacked Dataset

USGS
Gage Height Error Correction Model

Simulated model error

Time series of linear model errors

R² = 0.72
Gage Height Predictions

Final gage height prediction is a summation of the linear and error models.

Final gage height Predictions: $R^2$ 0.90 – 0.98
Simulation of Specific Conductance and Total Phosphorus

Two stage models:
- Static model using X, Y and measured data
- Dynamic model predict variability about mean

USGS
Decision Support System

- Excel application
- Integrates
  - Historical database
  - ANN and regression models
- Model controls
- Streaming graphics
- 3D visualization
- Model simulation output

Lox Empirical Model

[Image of Lox Empirical Model]

USGS

[Image of USGS logo]
Excel Spreadsheet

- Sheets:
  - Flow set points
  - Model simulation controls
  - Graphs
  - User define flow input
  - Tabular output
  - Release notes
  - Database
DSS Application: Canal water intrusion into the marsh

Intrusion events: Canal WL > Marsh WL
Negative slope

Blue = measured
Red = predicted
R²=0.87
Scenario 1

- What will be the model response to the simulated change in slope if the flow of Q4 (S-10D, S-10C, S-10A, and S-39) is increased by 40 percent?

Increase flow by 40%
DSS Application Set-up

- On Flow Set Point sheet\(^1\) (Q SPs) set flows for the Q4 structures to \(140\%\) of historical flows

Flow input options:
- \% historical flow
- Constant flow

All flows are set to \% of historical \(100\% = \text{actual flows}\)
DSS Simulation Controls

- Go to “Controls” sheet
- Set simulation period
- Write output
- Run Simulation
DSS Scenario Results

40% increase flow Q4

- **Blue** – Simulated actual slope
- **Red** – Simulated scenario – increase flow Q4 by 140%

Increasing the flows did increase the slope.
Negative slopes overall were minimized
Positive slopes also increased.
Scenario 2
Inflow = Outflow

Set Outflow = Inflows
DSS Scenario Results

Outflow = Inflow

- Blue – simulated actual slope
- Red – simulated scenario slope Outflow = Inflow

- Increasing the outflows did increase the slope.
- Negative slopes overall were minimized
- Positive slopes also increased.
Summary

- Model allows users to evaluate effects of flow releases
- Evaluate short- and long-term flow regimes
- Excel platform for DSS – facilitates dissemination of models user of various technical levels
- DSS database easily updated
- USGS report in final stage of review/production process.
Questions

Paul Conrads
pconrads@usgs.gov
803 750-6140

That's the whole problem with science. You've got a bunch of empiricists trying to describe things of unimaginable wonder.